DOCUMENT RESUME

ED 160 405 SE 024 973

AUTHOR	Helgeson, Stanley L., Ed.; Blosser, Patricia E., Ed.
TITLE	Investigations in Science Education, Vol. 3, No. 3. Expanded Abstracts and Critical Analyses of Recent Research.
INSTITUTION	Ohio State Univ., Columbus. Center for Science and Mathematics Education.
PUB DATE	77
NOTE	65p.
AVAILABLE FROM	Information Reference Center (ERIC/IRC), The Ohio State Univ., 1200 Chambers Rd., 3rd Floor, Columbus, OH 43212 (subscription $6.00, $1.75 single copy)
EDRS PRICE	MF-$0.83 HC-$3.50 Plus Postage.
DESCRIPTORS	*Abstracts; *Autoinstructional Methods; College Science; Curriculum Development; Educational Research; *Evaluation; *Instructional Design; *Science Education; Scientific Concepts; Secondary School Science; *Student Characteristics; Tests
IDENTIFIERS	Research Reports

ABSTRACT

This issue contains expanded abstracts of research reports grouped into two clusters and a section of individual studies. The first cluster contains abstracts of two research reports dealing with trait-treatment interaction studies. The second cluster deals with examination items categorized according to Bloom's Taxonomy. The final section, individual studies, includes four studies on the following topics: (1) student evaluation of instructors; (2) science misconceptions; (3) instructional design; and (4) self-directedness. Each abstract includes purpose, research rationale, research design and procedure, findings, interpretations, and an abstractor's analysis of the research. (GA)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED160405

INVESTIGATIONS IN SCIENCE EDUCATION

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Editor

Stanley L. Helgeson
The Ohio State University

Associate Editor

Patricia E. Blosser
The Ohio State University

Advisory Board

David P. Butts (1978)
University of Georgia

Kenneth G. Jacknicke (1978)
University of Alberta

Donald E. Reichard (1979)
Emory University

Ronald D. Anderson (1981)
University of Colorado

Frances Lawrenz (1980)
Minneapolis, Minnesota

Joe C. Long (1981)
University of Georgia

National Association for Research in Science Teaching

ERIC Clearinghouse for Science, Mathematics, and Environmental Education

Published Quarterly by

The Center for Science and Mathematics Education
College of Education
The Ohio State University
1945 North High Street
Columbus, Ohio 43210

Subscription Price: $6.00 per year. Single Copy Price: $1.75.
Add 25¢ for Canadian mailings and 50¢ for foreign mailings.

SE 024 973

NOTES . . .

from the Editor

This volume contains research reports grouped into two clusters and a section of individual studies. The first cluster, TRAIT-TREATMENT, contains analyses of two research reports dealing with trait-treatment interaction studies. The second cluster, BLOOM'S TAXONOMY, includes analyses of studies dealing with examination items categorized according to Bloom's Taxonomy. The final section, INDIVIDUAL STUDIES, includes four studies of such diverse topics as student evaluation of instructors, science misconceptions, instructional design, and self-directedness.

Publishable responses to these analyses or comments and suggestions for improvement of the journal are invited.

Stanley L. Helgeson
Editor

Patricia E. Blosser
Associate Editor

Volume 3, Number 3, 1977

NOTES from the Editor iii

TRAIT - TREATMENT 1

Ott, Mary Diederich. "Evaluation of Methods of Instruction and Procedures for Assigning Students to Methods." American Journal of Physics, 44 (1): 12-17, 1976
Abstracted by RONALD D. SIMPSON 3

Ott, Mary Diederich and David B. Macklin. "A Trait-Treatment Interaction in a College Physics Course." Journal of Research in Science Teaching, 12(2): 111-119, 1975.
Abstracted by DOROTHY GABEL 14

BLOOM'S TAXONOMY 19

Billeh, Victor Y. "An Analysis of Teacher-Made Science Test Items in Light of the Taxonomic Objectives of Education." Science Education, 58(3): 313-319, 1974.
Abstracted by DAVID H.OST. 21

White, Richard T. and Lindsay D. Mackay. "Does Bloom's Taxonomy Apply to Physics Examinations?" The Australian Science Teachers Journal, 18(4): 66-70, 1972.
Abstracted by ROBERT L. STEINER. 27

INDIVIDUAL STUDIES. 33

Levenson, Hanna and David W. Brooks. "Student Evaluation of Lectures versus Graduate Student Laboratory Instructors in Introductory College Chemistry." Journal of College Science Teaching, 5(2): 85-88, 1975.
Abstracted by RICHARD M. SCHLENKER 35

Za'rour, George I. "Science Misconceptions Among Certain Groups of Students in Lebanon." Journal of Research in Science Teaching, 12(4): 385-391, 1975.
Abstracted by EUGENE D. GENNARO 44

Burkman, Ernest. "An Approach to Instructional Design for Massive Classroom Impact." Journal of Research in Science Teaching, 11(1): 53-59, 1974.
Abstracted by ROBERT E. YAGER 52

McCurdy, D. W. "An Analysis of Qualities of Self-directedness as Related to Selected Characteristics of I.S.C.S. Students." Science Education, 59(1): 5-12, 1975.
Abstracted by GERALD G. NEUFELD. 59

T R A I T - T R E A T M E N T

Ott, Mary Diederich. "Evaluation of Methods of Instruction and Procedures for Assigning Students to Methods." American Journal of Physics, 44 (1):12-17, 1976.

Descriptors--*Autoinstructional Methods; *Achievement; *Attitudes; College Science; Educational Research; Evaluation; Higher Education; *Instruction; *Physics; Science Education

Expanded Abstract and Analysis Prepared Especially for I.S.E. by Ronald D. Simpson, North Carolina State University.

Purpose

The general purpose of this investigation was twofold. It was designed to evaluate two methods of instruction in a physics course at Cornell University. The study also examined two procedures for assigning students to these methods.

The two methods of instruction compared were audio-tutorial (AT) and lecture-recitation-laboratory (standard). The two procedures for placement of students were random assignment and assignment according to student preference.

The investigators also sought to compare immediate and longer-term effects related to both achievement and attitude among the four treatment groups. Data from the 1974 study were compared to data from a similar investigation conducted in 1973.

Rationale

Physics 112, Mechanics and Heat, is a one-semester introductory course offered to approximately 550 engineering and physics majors each spring at Cornell University. Most of the students are freshmen level males. Faculty working with this course were interested in comparing achievement and attitudes between two methods on instruction: AT and standard. Their interest, however, went beyond merely seeking to determine which method was superior. They were also interested in learning which method was more suitable for certain students. They were interested in comparing

achievement and attitudes between students randomly assigned to the two methods and those allowed to enroll according to preference. Other variables such as Scholastic Aptitude Test (SAT) scores, mathematics achievement pre-test scores, and attendance records were compared across groups in what the investigators termed a "trait-treatment-interaction" approach.

In short, the rationale for this study was to learn if different types of students at Cornell University performed in a differential manner when studying introductory physics via two contrasting methods. Likewise, there was interest in learning if an interaction existed between student attitudes and the instructional methods used. Furthermore, there was interest in exploring other student traits in light of their cognitive and affective behavior.

Research Design and Procedure

At Cornell University approximately 270 students taking Physics 112, Mechanics and Heat, were assigned to four treatment groups during the spring semester of 1974. The treatment groups related to the methods of instruction and to the procedures of assigning students to the methods. Hence, the four groups were: audio-tutorial-random (ATR), audio-tutorial-preference (ATP), standard-random (STR), and standard-preference (STP). Students in all groups had the same textual materials, the same homework and laboratory assignments, identical quizzes and examinations, and roughly the same content. The standard method included two hours of lecture and two hours of recitation every week and a two-hour laboratory period every other week. The AT method included one hour of recitation per week and was designed primarily to allow group interaction as well as student contact with one particular instructor. All other instruction with this method took place at the student's convenience in a learning center staffed by tutors 47 hours per week. Materials in the learning center included apparatus for self-demonstrations, the same selection of laboratory equipment available in the standard laboratories, audio-tape commentaries and slides. These materials emphasized concept development as well as problem solving and were used in addition to the course textbook and supplementary notes.

Students in this study were assigned to 1 of 15 sections scheduled during four different class hours. (Students were enrolled in other sections but were not used in this study.) Eight of these sections became the standard group and seven became AT recitation sections. At two of the four class hours, students were allowed to choose the treatment based on their preference of teaching methods. As a result of further random assignment, there resulted four ATR sections, three ATP sections and four sections each of STR and STP. The 15 recitation sections in this study were taught by 8 teachers, including 2 faculty members and 6 graduate students. Seven of the teachers taught one AT and one standard section each. One graduate student taught only one standard section included in this study.

To investigate possible differences between the effects of the methods on different student groups, the Johnson-Neyman technique of regression analysis was used. The author stated that regression analysis was used instead of analysis of variance because the latter reduces continuous scores to a small number of levels, making this procedure inefficient.

The basic measure of student achievement in this study was final grades. This variable was based on recitation performance, laboratory reports, two preliminary examinations and a final examination. The "traits" analyzed in this study were achievement in mathematics as measured by a course-specific pre-test and mathematics aptitude as measured by students' scores on the mathematics portion of the SAT.

Student attitudes toward the AT and standard methods were measured by responses to a questionnaire completed by students at their last recitation meeting of the semester. Two "fairly global statements" were included in each questionnaire: "In general, I have been satisfied with the AT (standard) method of instruction used in P112" and "I am glad I took the AT (standard) version of P112 rather than the standard (AT) version."

Longer range effects of the differences in methods of instruction were considered in this study. This was accomplished by analyzing the progress of students in a similar 1973 study. The investigators sought to

determine (1) whether students from the four treatment groups enrolled in the same types of engineering courses, (2) if there were any negative effects of the AT method when students returned to a lecture-recitation-lab format in their science courses, and (3) whether AT and standard students had different attrition rates in later semesters.

Findings

The design of this study (1974) was based, at least in part, on results by the same investigators from a similar study conducted during 1973. The results of the earlier research indicated that an interaction existed between two student "traits" (mathematics achievement as measured by a course-specific pre-test and mathematics aptitude as measured by the SAT) and the two methods of instruction. Students with very high mathematics aptitude (SAT math scores of 725 or higher) and high mathematics achievement on the pretest had higher predicted grades in the standard method than did comparable students in the AT method. Students with relatively low mathematics aptitude (SAT math scores of 625 or lower) and low mathematics achievement had higher predicted grades in the AT method than did their counterparts in the standard method. Using the same linear regression techniques, predicted grades of students ranking intermediate in mathematics aptitude and achievement did not differ significantly in the two methods. Additionally, in the 1973 study there were differences in course grades within the AT method between the randomly assigned students and the students assigned by preference. When mathematics aptitude and achievement were controlled, AT students who had been assigned randomly had significantly higher achievement than did those assigned by preference. The 1974 investigation was, therefore, designed in part to determine whether these 1973 findings would be reproduced. To this end, previous results were not revealed to course instructors of the 1974 study and attempts were made to maintain similar content coverage across years of treatments.

Results from the 1974 investigation demonstrated no significant difference between the regression lines of the AT and standard groups when predicted grades were related to mathematics pretest and aptitude scores in an overall comparison between methods.

Major findings from the 1974 investigation revealed that student achievement difference existed between the random and preference assignment procedures. Among students who were randomly assigned the methods, mathematics achievement was significantly related to grades in each method of instruction. The slopes of the regression lines relating pretest mathematics achievement to predicted physics grades did not differ significantly in the two methods. Mathematics aptitude as measured by the SAT, conversely, was not highly related to grade in either method for randomly assigned students. Investigators, therefore, concluded that there was no evidence of a trait-treatment-interaction among the randomly assigned students. Among students assigned by preference, however, differences did exist. Those who selected the standard method achieved higher final grades than did those who selected AT. When predicted grades were related to both mathematics achievement and aptitude, the predicted grade was higher in every case for the STP group (over the ATP group) when the Johnson-Neyman technique was used. In particular, for a range of SAT mathematics values of about 630-670, the Johnson-Neyman analysis predicted significantly higher (at 0.05) grades in the STP group.

Student attendance patterns were different between the ATP and STP groups. Seven of 45 ATP students responding to a final questionnaire indicated they stopped attending the learning center on a regular basis by the fourth week of the course. None of the 74 STP students indicated on this questionnaire that they had stopped attending lecture that early in the semester. In addition, 24 of the ATP students said they did not regularly attend the learning center. STP students reported spending 8.6 hours per week outside of class while ATP students reported 5. STP students exhibited a considerably higher average grade on the first exam than ATP students while grades on other tests did not differ significantly. ATP students also possessed slightly lower grade point averages in their other courses when compared to STP students.

When students in the 1973 and 1974 studies responded to two Likert-type attitude statements ("In general, I have been satisfied with the AT/standard method of instruction used in P112" and "I am glad I took the AT/standard version of P112 rather than the standard/AT version."), the

following results were obtained: (1) students assigned by preference expressed more positive attitudes toward the method they received than did students assigned randomly; (2) AT and standard students tended to be equally satisfied with the method they received; and (3) standard students, in 1974, but not in 1973, were more likely than AT students to say that they were glad to have taken their method of instruction rather than the other method. The author has been quick to point out, however, that increased disaffection with AT could well have been due to problems associated with "coordinating the two methods of instruction."

Follow-up behavior of approximately 90 percent of the students in the 1973 study was analyzed. AT and standard students enrolled in roughly the same selection of engineering courses subsequent to the 1973 investigation. Considering students' grades in sophomore engineering and physics courses, ATR performed as well as, and in some cases better than, STR students. STP students, on the other hand, performed somewhat better in some courses than their ATR counterparts. While students in ATR outperformed those in STR and STP outperformed ATP students in other courses, when taken together there were no apparent overall differences in achievement between the two teaching methods. Comparing attrition rates among the four groups produced a similar pattern—that of attrition being inversely related to achievement. Again, however, when the two methods were compared overall there appeared to be no significant difference between standard and AT instructional groups.

Interpretations

One of the purposes of this study was to compare student achievement of students at Cornell University enrolled in a physics course taught by two contrasting methods of instruction: audio-tutorial and standard lecture-recitation-laboratory. The most significant finding of this investigation was the fact that when students were assigned to the two teaching methods by random procedures, there was no difference in achievement between treatment groups. When students were allowed to select the instructional method based on preference, however, the group taught by the standard method exhibited superior achievement in physics.

Considering attitudes alongside achievement, results from this study indicated that students assigned by preference were more satisfied with the method they received than were the randomly assigned students (although, AT students were more likely than standard students to say they would have preferred the alternative method). Follow-up studies did not indicate any appreciable differences in enrollment patterns, course grades, or attrition rates between students taught by the two methods.

The aforementioned summary suggested to the investigators of the report a dilemma. If students are allowed to choose between two methods of instruction such as lecture-recitation-laboratory and audio-tutorial, they are allowed a greater degree of flexibility and self-determination. Results from this study suggest that this leads to greater student satisfaction, to more positive attitudes toward the method of instruction in which students are engaged. In the case of the students in this study who were allowed to take AT physics because they preferred it, however, a lesser degree of achievement ensued. One reason offered for this was that the students were apparently overconfident. These students as a group had taken more physics in high school and exhibited somewhat higher mathematics aptitude as measured by the SAT. Additional data showed that they spent less time studying out of class and that they were more likely to stop attending the learning center regularly. The investigators offered no solution to the dilemma. They did suggest, though, that additional indicators of student achievement and attitude need to be explored in relation to teaching effectiveness. Gains in such considerations as student independence, self-confidence, interest in the subject matter and desire to take additional physics courses would certainly represent factors that might tend to mediate the results produced in this study.

ABSTRACTOR'S ANALYSIS

One of the most frequently asked questions of college level instructors is "Which teaching method, or methods, should I use in my classroom?" Among the most frequently discussed alternative teaching strategies in

higher education is the auto- or audio-tutorial (AT) method. Some display of this approach or modification thereof can be found somewhere within most colleges and universities. One of the most tempting areas of educational research is that of comparing student performance among a variety of teaching approaches. Since AT represents a rather dramatic and visible shift in American pedagogy, it is only natural that college level educators would be interested in comparing it with traditional methods. The literature has indeed contained many such studies over the last decade. Yet, in most cases, results have been cloudy and inconclusive. And, when results from one study are compared with those of another, it is usually difficult to formulate substantial generalizations.

One of the reasons it is difficult to draw conclusions from methodological studies of the type just described is that within any investigative setting exist variables which are next to impossible to duplicate. This will likely remain a drawback to educational research of this type. But, there is another traditional weakness of studies of this type and it is that the dependent variable generally consists of a single measure—that of factual recall of material on standardized or teacher-made tests. The potentially powerful effects of an alternative teaching method is often never measured when a single measurement of achievement is used. There is at least a third reason why educational studies of this type are often inconclusive or meaningless. They do not consider the nature of the student. They frequently fail to consider the fact that individual students react very differently to the same stimuli, simply because each student has different needs, different attitudes, different cognitive styles and different levels of maturity.

When considered within the matrix of other investigations of this type, this study represents an advancement in both knowledge and research methodology. New relationships have been forwarded here with respect to how additional student variables or "traits" may mediate achievement and attitude. Furthermore, the use of such techniques as preliminary, follow-up, short-range and longer-range measures, in both cognitive and affective dimensions, represent potential parameters in which new relationships can be discovered. Had this study been involved with merely assigning students to two treatment groups and comparing course

achievement, no significant differences would likely have existed. But the research design in this study allowed for the construct "preference" to be considered; and, as a result we learned that there is an apparent difference in the way engineering students at Cornell University perform in Mechanics and Heat when they are allowed to choose between two teaching methods. Likewise, in measuring such variables as attitude toward the AT method of instruction, attitude toward the lecture method, attendance, attrition, hours of study per week, subsequent selection of courses, number of physics courses in high school, mathematics aptitude, and other pertinent variables, it is possible to compare the two teaching methods within more than one context. In essence, this study avoids a trap of oversimplification. Instead of being forced to choose between two teaching methods—to declare one the "best"—educators and researchers are free to examine several relationships that emerge from this investigation. Further, there is something concrete on which to build other studies.

Would an investigation of this design produce similar results at other institutions? Since engineering and physics students, Mechanics and Heat, and faculty personnel at Cornell University are no doubt unique to other settings, the external validity of this study has not been established. Reasonable safeguards were taken, however, to build a case for internal validity. The regression analysis procedure used in this study appears to be an appropriate statistic. The author of this paper has been quick to point out potential weaknesses and concerns. One such factor related to the heterogeneity of the student population in the 1974 study. Apparently, many higher ability students took Physics 112 during the fall of 1973, leaving perhaps a skewed population during the spring semester of 1974. In order to establish external validity this study should be repeated at other institutions with other student populations.

I found this report well-written and easy to follow. It presented a rationale that was defined clearly and logically. The sample and procedures were documented and pertinent data were displayed. In practically all cases, information in the tables and figures was developed adequately in the body of the paper. The results and conclusions were stated

carefully and congruently. The reasoning used to develop relationships among the variables was easy to follow and understand.

Research of this type is always difficult and it is not always possible to control all the variables to the extent one would prefer. As stated earlier, the author cited several things beyond her control in this research design. Perhaps the most noticeable lack of control came when a change in professor decreased the degree of coordination in the content of the two methods and between the 1973 and 1974 studies. In fact, the variable "professor" or "professors" constitutes a powerful treatment effect in studies of this type and, when not controlled, can lead to some hard-to-answer questions. The "like" or "dislike" by students of a key faculty member or two in a study like this (especially one with a small N) could mask attitudes toward "teaching method." In a couple other instances, it was difficult to see whether data from the 1973 study could be rightfully compared with data from the 1974 study.

My biggest criticism of the study is that so few affective measures were taken. It is quite possible that students were focusing their feelings on objects of affect other than "teaching method." Attitudes toward "physics," "teacher," "college," "engineering," "studying," "grades," and "academic self" represent areas that could be more important to freshman level students than "AT" or "lecture." Of course, the author alluded to this at the end of the report. She also mentioned variables such as student independence and self-confidence, which indeed could be significantly influenced by a teaching method such as AT which places more of the responsibility for learning on the student. While students often insist that they prefer "flexibility" and "freedom," it is not surprising that they might not only perform better but actually feel more secure with a traditional teaching method, one under which they had been previously nurtured.

I think this investigation can serve as a model on which to build further studies of this type. By introducing and controlling for additional cognitive and affective variables, additional relationships can be sought. By using regression analysis it is possible to develop research designs

that allow researchers to examine crucial factors as both independent and dependent variables. For example, variables such as "mathematics achievement" and "academic self-concept" are both entry traits as well as exiting traits. These and other variables not only influence achievement but are influenced by achievement. Students possess these traits when they come to us and they possess them when they leave us. Often the effect of one course is not powerful enough to induce significant change. Studies of this type should include long-range components. For instance, the effects of one AT course may not be significant, but several courses where students must learn to be more independent may cause a change that could be measured later. It is important to remember that one course or one professor or one teaching method is but a small portion of the total educational experience of a student.

As more studies like this one are conducted, educators will slowly uncover additional relationships central to the teaching-learning process. As this happens, we will surely be able to improve the instructional setting and academic performance of our students. This study is an example of how a carefully planned, carefully conducted, and well-reported investigation can help expand the body of knowledge in college teaching.

Ott, Mary Diederich and David B. Macklin. "A Trait-Treatment Interaction in a College Physics Course." Journal of Research in Science Teaching, 12(2): 111-119, 1975.

Descriptors--*Autoinstructional Methods; *College Science; Educational Research; Higher Education; Instruction; *Physics; Science Education; *Student Characteristics; *Teaching Methods.

Expanded Abstract and Analysis Prepared Especially for I.S.E. by Dorothy Gabel, University of Indiana.

Purpose

The purpose of this study was not to determine if audio-tutorial instruction was superior to conventional instruction but to determine if an interaction between the type of instruction (audio-tutorial or conventional) and students' traits (mathematical ability) had an effect on achievement.

Rationale

Frequently no significant differences are reported in student achievement due to differences in instructional strategies. In this study the authors follow the suggestions of Berlinger and Cahen (1973) and Cronbach and Snow (1969) to examine interaction between student traits and instructional methods. Such findings should enable educators and students to select instructional strategies that enhance particular students' achievement.

Research Design and Procedure

The sample consisted of 303 college engineering and physics students (98 percent freshmen) enrolled in a one semester introductory level physics course. Students who had a scheduled recitation period at a given hour were randomly assigned to a class section which had either audio-tutorial or traditional instruction. In about half the cases,

however, students were allowed to indicate a preference for a particular type of instruction. In these cases the students were randomly assigned to one of the sections with their preferred model of instruction. This resulted in 115 in the audio-tutorial treatment (57 randomly assigned and 58 selecting it) and 188 in the standard treatment (101 random, 87 preferred) in each. These students were distributed in 15 recitation sections.

The treatment consisted of instruction by two different modes, audio-tutorial and conventional. The audio-tutorial method included one hour of recitation per week. All other instruction took place at the student's convenience in a learning center staffed by tutors 52 hours per week. The standard or conventional method included two hours of lecture and two hours of recitation per week and a two hour lab every other week.

Course content in both treatments was comparable. Both groups had the same homework assignments, similar lab experiemtns, and the same examinations. The lecturer of the standard course was the co-author of the materials used in the audio-tutorial course.

In order to determine students' original traits, students were given a questionnaire and a math pretest at the first recitation period. The authors then selected traits on the basis of (1) the importance of the trait to the course, (2) the reliability and validity of the measurement of the trait, and (3) range of responses to the trait measure. Three traits met the criteria: the College Entrance Examination Board Scholastic Aptitude Test in Mathematics (SATM), verbal aptitude of the same test (SATV) and a mathematics pretest composed by the authors. SATV was eliminated because it was found not significantly related to the final grade in either treatment.

The dependent variable, student's achievement in the course, was then measured by the final course grade. This was based on the student's lab work, quizzes, interim examination and final examination.

data were analyzed in several ways. Means and standard deviations of the SATM, Math Pretest and Final Grade were compared for the two methods of instruction using a t-test (not explicitly stated). Regression equations were formulated to predict estimated grades resulting from the two instructional modes. The Johnson and Neyman technique (1936) was used to analyze the interaction between the trait (mathematical ability) and treatment (method of instruction).

Findings

Although there were some significant differences in pretest math scores and SATM variances at the 0.05 level, there were no significant differences in the means or variances of final grades in the course.

The following regression equations were formulated:

Audio-Tutorial $Z = 0.5175X_1 + 0.2633Y_1 + 9.1667$

Standard $A = 0.8495X_2 + 0.4081Y_2 + 10.6352$

where Z and A are course grades, X is the score on the math pretest and Y the level on SATM.

The Johnson and Neyman technique indicated that at the 0.10 level of significance the standard treatment was preferable to the audio-tutorial in terms of expected achievement for students with math pretest scores equal to 9 and SATM greater than or equal to 725. The audio-tutorial treatment was preferable for students with math pretest scores less than or equal to 4 and SATM less than 625. Both methods of instruction were equivalent for all other students.

Interpretations

Results of this experiment indicated that although there was no significant difference in physics achievement for students who studied using the audio-tutorial method of standard methods, there was a significant interaction

between physics achievement and math skills as measured by a pretest and the SATM. Students with high ability profit more by the conventional treatment whereas students of less ability profit from the audio-tutorial method. Because this study is limited to a specific physics course, results cannot be generalized. Similar studies should be conducted to determine if this trait-treatment interaction is applicable in other courses.

ABSTRACTOR'S ANALYSIS

Although this study has limited generalizability, great care was taken by the authors to randomly assign students to treatments and to control mediating and extraneous variables. In addition, the study is concerned with an area of research, trait-treatment interaction, that is of growing interest and concern to science education.

There are several areas in which this research and the report could be strengthened. First, the purpose of dividing students into two groups, those that preferred a particular mode of instruction and a randomly assigned group, is not clear. The authors do not justify this classification in their rationale, do not state hypotheses concerning it, and do not make sufficient use in their analysis. These data probably could be further analyzed to compare differences in achievement between students who selected audio-tutorial instruction and those who were assigned to it.

A second concern is with the validity and reliability of the instruments used to measure the students' mathematical skills. One of the two instruments used is a nine-item multiple choice math pretest. The authors state that the correlation between this test and the SATM test they administered was 0.33 and therefore appeared to measure different traits. Although this may be true, no mention is made about the reliability of this test. A nine-item test may have low reliability and therefore invalidate the study.

Another area that needs clarification in the article is in the interpretation of the tables. Although the number of subjects in Table III is listed as 303 students, the number in Table II lists 282 students. It is also difficult to interpret the means in Table III. The dependent variable is listed as the course final grade based on scores that range from 0 to 400. The reader really needs to know how these scores are translated into grades. A final grade of 6.44 has little meaning as one does not know to what letter grade it is equivalent or how it was derived. Because of this, it is also difficult to interpret the regression equations that are given for both treatments.

In addition, specific mention should be made of which statistical tests are being used. One could probably assume that a t-test is used for the analysis of the means in Table III but no mention of which correlation coefficient was used is given in the article.

A contribution that this study makes to this area of research is in the area of methodology. The methodology for determining the trait-treatment interaction appears sound. The method used was the Johnson-Neyman technique. By using this technique one cannot only determine whether there is a significant interaction but also the level of the trait that will yield a significant interaction at the level of one's choice. Other researchers who investigate trait-treatment interaction may wish to examine this methodology to determine its applicability to their own studies.

Studies of this nature make a significant contribution to science education even though the results may not be generalized to other courses. This study acts as a model on which replication studies in other educational settings can be carefully conducted. By combining results from a series of these replications, generalizations can be made on successful instructional strategies for various types of students in diverse settings.

REFERENCES

Berliner, D. C. and L. S. Cahen. Review of Research in Education, 1:58,1973.

Cronbach, L.J. and R.E. Snow. "Individual Differences in Learning Ability as a Function of Instructional Variables." Stanford University,pp.1-25,1969.

Johnson, P. O. and J. Neyman. Statistical Research Memoirs, 57:1, 1936.

BLOOM'S TAXONOMY

Billeh, Victor Y. "An Analysis of Teacher-Made Science Test Items in Light of the Taxonomic Objectives of Education." Science Education, 58(3):313-319, 1974.

Descriptors--Achievement Tests; Educational Research; *Evaluation; Instruction; *Secondary School Science; Teacher Characteristics; Test Construction; *Tests

Expanded Abstract and Analysis Prepared Especially for I.S.E. by David H. Ost, California State College, Bakersfield.

Purpose

The study is described as having two major purposes: (1) to identify the pattern* of the cognitive processes implied in teacher-made examinations in secondary school science; and, (2) to relate the pattern to selected variables (i.e., grade level, subject matter taught, and some teacher characteristics). From these two major purposes, four questions were asked:

1. What relationship, if any, exists between the level of the test item as classified by Bloom's Taxonomy and each of the following teacher characteristics: years of science teaching experience, professional in-service training, academic specialization, and status as part-time versus full-time?
2. Is there a relationship between the science subject taught and the level of the test item?
3. Do teachers of different grade levels emphasize different cognitive levels of test items?
4. What percent (proportion) of the test items asked by secondary school science teachers falls into each of the categories of Bloom's classification system?

Rationale

The relatively recent attention given to the types and strategies of asking questions has yielded considerable insight into the complexities

---

*Billeh does not explain what he means by a pattern. He does not appear to discuss the results of his research in terms of patterns.

of the process. For example, it has been shown that through proper questions, the teacher can have a significant influence on directing and developing the cognitive processes of students. Test items are questions and must be included in these types of studies. Questioning, in the form of examinations, is subject to the same forms of analysis. It is therefore necessary that questions be classified in a manner which identifies the level of cognitive process necessary to develop an appropriate answer to the question. Billeh suggests that in determining the cognitive level of a particular test item, the judgment must be made in the context of the learning experiences, instructional materials, and other factors to which the students were exposed or which were used to stimulate learning of the material represented by the test item.

Research Design and Procedure

Twenty-five randomly selected secondary schools in Beirut, Lebanon, which had at least grades seven to ten were identified for the study. One class each of grades seven and ten was selected in each school.

Tape recorders were used to obtain a verbatim record of three-to-five hour sessions covering one science unit in each class. Each teacher participating in the study was asked to develop an hour examination of the unit which was recorded. Thirty-three examinations were submitted. In addition, information about the teachers' academic specialization, teaching experience, training, and status was obtained.

Three jurors experienced in Bloom's taxonomy worked independently on classifying the test items according to Bloom's classification system. Consensus was later reached as to the cognitive level of each test item. The percent of test items falling into each level was calculated on using the weight assigned to each item by the teacher. The mean proportion of test items representing each cognitive level was calculated.

Use of one-way analysis of variance was made to analyze the relationship between the science subject taught and the level of the test item. The t-test was used to determine whether cognitive levels differed between grades seven and ten.

Pearson's correlation coefficient was computed for the relationship of teaching experience and the proportion of test items at each level. Similarly, the biserial correlation coefficient was computed and tested for significance to determine whether the teacher variables were related to the proportion of test items in each cognitive level.

Findings

The statistical results showed that no relationship existed between professional inservice training and cognitive levels. A "moderate relationship" was found between the teacher's field of specialization and the level of test items, with those teachers trained in science asking questions that require comprehension and application. A "low correlation" existed between the teacher's status and cognitive level questions. Full-time teachers required somewhat higher levels of cognition.

The results of the analysis of variance indicated that no relationship "seems to exist" between the level of the test item and the subject matter taught. Similarly, there were no significant differences found for the levels of items generated by teachers of grade seven in comparison with teachers of grade ten.

The Pearson product moment correlation analysis of the relationship between teaching experience and the proportion of items in each cognitive level showed a "moderate positive relationship" between the knowledge category and teaching experience (reportedly significant at the 0.01 level) and a "low negative relationship" when the correlation between teaching experience and the items classified as Comprehension and Application.

The following is quoted from the article:

a. by far, the heaviest emphasis in science examinations is on the lowest level of the classification system--the knowledge level. Nearly 72 percent of examination time is devoted to recall of facts...60 percent of that time requires only the lowest cognitive levels;

b. only 7 percent of the examination's time is devoted to questions requiring the application of science principles, theories, or other abstractions to new situations;

c. test items requiring comprehension constitute 21 percent of the examinations;

d. test items requiring the highest cognitive levels, namely analysis, synthesis, and evaluation are absent.

Interpretations

Billeh concludes that the examination items prepared by the science teachers in Beirut are in the lowest subcategories of the knowledge category; there are no effective "critical thinking" items. He attributes this to inadequate training in testing and evaluation and the inability to identify important educational objectives. Billeh suggests that objective identification is a necessary prerequisite to the development of effective questions; such test items may be useful in the development of students' cognitive processes.

ABSTRACTOR'S ANALYSIS

Unlike many educational research efforts, the significance of the results lies in the fact that significant differences between the various populations were not found. It would seem logical to hypothesize differing levels of questions when comparing seventh and tenth grade teachers, better trained and lesser trained teachers, or the different science courses taught. Such differences were not found. Perhaps the most startling finding reported was that more items requiring low levels of cognition seem to be asked by teachers with the greater amount of teaching experience. This would appear to be worthy of further study.

In general, the study contributes to the body of knowledge centering on questioning, objective formation, and evaluation. There are however, several items in this report that must be discussed, some are minor technical questions, others are somewhat broader in scope.

Although the author indicates that the schools involved in the study were selected at random, he does not indicate how the specific teachers (classes) were selected in those schools. If 25 schools were involved in the study and each school contained two classes, that would indicate that 50 teachers would be involved. Billeh reports that the 33 examinations used in the study were submitted by teachers in 18 schools. What happened to the other seven schools and 17 teachers? Was there some selection that may have biased the results?

Another component of the study that is not thoroughly discussed is the process of selection of the three-to-five hour sessions covering the unit in each class. How was the unit selected? Was there randomizations or an attempt to maintain similarity? It would seem that in some cases that to develop a one-hour exam covering material that was learned in a three- to five-hour session may force the teacher to over-emphasize low level cognitive skills. As Billeh points out, there was not even one question at the analysis or synthesis level. Somehow, the design should insure that both the type and amount of content covered would be both of quality and quantity to provide the background necessary for the developing of higher cognitive questions. If the three-hour class session is devoted to low level cognitive learning, which may be very appropriate in the context of the overall curriculum, why would the examination be any different in level?

The establishment of the alpha levels ranging from 0.01 to 0.05 was done so without any commentary. It appears that the levels of statistical significance were used more for descriptive purposes than in terms of statistical inference. If calculations are being used for descriptive purposes, then the level of significance is of less value than a discussion of Type 1 and Type 2 errors associated with the study. The researcher's comments concerning Type 1 and Type 2 errors would be of as much, if not more, value to the reader than are the footnotes, "Not

significant at 0.05 level," or "significantly different from zero at the 0.01 level." Although this abstractor recognizes the problem faced by authors in getting research published that does not include some aspect or tests of statistical significance, there would be considerable benefit accrued to the research community if some space and time were devoted to the items described above. To be of maximum use, descriptive studies must provide as much of the background data and milieu of the investigation as possible.

The investigator clearly limits the conclusions drawn from the study to science teachers in Lebanon. Since the sample was drawn from that population, it seems reasonable however, that there are several hypotheses which are inherent in those conclusions that require further investigation. Perhaps the most important is that "science teachers need to be trained in identifying important educational objectives in a specific teaching setting." While this may seem logical, such a statement does not necessarily follow from the results provided in this research paper. An equally valid hypothesis, recognized by Billeh, is that the real issue may lie in the teachers' inability to develop test items of a higher cognitive level. Perhaps these two issues go hand-in-hand, yet that is an assumption that must also be tested.

White, Richard T. and Lindsay D. Mackay. "Does Bloom's Taxonomy Apply to Physics Examinations?" The Australian Science Teachers Journal, 18(4):66-70, 1972.

Descriptors--*Achievement Tests; Evaluation; Evaluation Criteria; *Educational Research; *Physics; Science Education; *Secondary School Science; Test Construction

Expanded Abstract and Analysis Prepared Especially for I.S.E. by Robert L. Steiner, The Ohio State University.

Purpose

The researchers investigated standardized examination item characteristics in terms of their possible importance in the preparation of valid examinations. The Victorian (Victoria, Australia) Matriculation Physics Examinations of 1966, 1967, 1968 were analyzed in terms of the subject matter content of the items, the item response format, and the cognitive level of the item , to see if these were relevant dimensions to be considered in the construction of valid examinations.

Rationale

The use of dimensional grids to assist in preparing course examinations is a common practice. It has generally been assumed that the validity of examinations is indeed improved if items are selected according to cognitive ability level. It is expected that items with similar characteristics should be highly related.

Research Design and Procedure

This correlational study was based on an analysis of representative samples of student data for the Victorian Matriculation Physics Examinations of 1966, 1967 and 1968. Each of the items contained in the examinations was classified on three dimensions. The dimensions were:

1. Subject Matter Content (3 divisions). The examination items were designed for the four sections of the PSSC physics textbook (prior to the 3rd edition) with Sections I and III forming one division, Section II a second division and Section IV the third division.

2. Item Response Format (2 divisions). The items were either of the completion or multiple choice response format and were classified accordingly.

3. Cognitive Level (2 divisions). Knowledge and comprehension items were classified in one division and all higher cognitive level items were grouped in a second division.

Each item was uniquely classified as fitting one of the twelve possible cells of the three dimensions.

Subject Matter Content

Cognitive Level		PSSC I & III		PSSC II		PSSC IV	
		Format		Format		Format	
		MC	C	MC	C	MC	C
	I						
	II						

Student response to each item was scored "1" if correct and "0" if incorrect. Student score for each of the 12 cells was the sum of correct responses to items classified in the particular cell. Twelve-point biserial correlations were made for each of the items contained on the 1966, 1967 and 1968 examinations utilizing a representative student sample for each examination. The 12-point biserial correlations for each item were made between student score on individual items and student total score on each of the 12 cells. In the case of the correlation of the item to the cell to which it was classified, the total score did not include the item being correlated.

The correlation matrix for each examination was used to carry out four separate analyses. In each of the analyses, the number of items correlating highest with cells of similar item characteristic dimensions as

compared to the number expected due to chance was used for statistical analysis.

In the first analysis, item correlations with all 12 cells were considered. The frequency of the highest correlation of items to cells to which they belonged was examined.

In the second analysis, the frequency of the highest correlation of items to the 11 cells other than the one to which they belonged was examined.

In the third analysis, the frequency of the highest correlation of the items to the eight cells of different content only was examined. This was carried out for the 1966 examination only.

In a fourth analysis, the frequency of the highest correlation of items to the four cells of different content and item response format only was made for all three examinations.

Findings

In the first analysis, it was found for all examinations that the number of items which correlated highest with the cell to which they were uniquely classified was significant ($p < .0001$).

In the second analysis, the cell to which the item was uniquely classified was restricted from the analysis and the item correlations to the remaining 11 cells was examined. The number of items in which the highest correlation was with a cell of similar content was significant for the 1966 and 1967 examinations ($p < .0001$) and the 1968 examination ($p < .01$). The number of items whose highest correlation was with cells of similar response format was also significant for the 1967 and 1968 examinations ($p < .02$). The number of items with highest correlation with cells of similar cognitive level was not significant.

In the third analysis for the 1966 examination only, correlations of items with cells of different content only indicated that the number of items

whose highest correlation was with cells of similar response format was significant ($p < .005$).

In the fourth analysis, which looked only at item correlations with cells of different content and response format, the number of items whose highest correlation was with cells of similar cognitive level was not found to be significant.

Interpretations

Based on the results of the four analyses, the researchers concluded that both content and response format were relevant dimensions to consider in items and should be considered in ensuring content validity in Grade 12 physics examinations. The researchers also concluded that there was no evidence from their analyses of the 1966, 1967 and 1968 Matriculation Physics Examinations of the cognitive levels of classification as used in the study to indicate that it was a relevant dimension to be considered in relation to content validity when examinations are constructed.

ABSTRACTOR'S ANALYSIS

The methods used to investigate content validity in this study are different from those traditionally used in evaluation studies. It is not apparent that the investigators have examined content validity, but rather the strength of relationships between individual items and post hoc a priori dimensions of physics examinations.

The results are not particularly surprising and perhaps predictable. The dominant role of subject matter content has been shown in many studies and one would expect the highest correlation to exist here. The authors found that the frequency of correlation of items to the cells of similar cognitive level was not different from that due to chance and concluded that cognitive level was not a significant dimension to consider in content validity of physics examinations. The use of only two levels of

cognitive ability should improve the accuracy and reliability of the item classification, but there are still some inherent problems with the classification. Students were coming from numerous backgrounds of schools and teachers. What may have been higher cognitive level items for some may not have been so for others. The other dimensions of content and item response format do not suffer from the possible overlap or mixing of subgroups as the cognitive level classification does.

The report is lacking somewhat in clarity. This could be due in part to the brevity of the article. Fuller explanations of some aspects of the study and the inclusion of some data and tables would have made the report much more meaningful to the reader.

The authors indicate that a representative student sample was used, but there is no indication of the sample size, or the size of the student population from which the sample was selected, or how it was selected. The authors also indicate that some of the cells to which the items were uniquely classified contained a small number of items, but there is no indication in the report as to how many items. A table indicating the number of items contained in each cell for each examination would have been beneficial.

The frequency of highest item correlation to cells of similar item characteristic dimensions compared to that based on chance was used as a basis for conclusions. Although probability levels are given, there is no indication of the actual test statistic used to determine significant results.

The magnitude of the correlations is not given nor indicated in the report. The analyses are based on the largest item-cell correlations, but there is no way of knowing if the correlations are statistically or educationally significant. Since only the largest of the item cell correlations was used in each analysis, other significant item-cell correlations would not be considered. Again a table(s) for each of the four analyses actually giving the number of correlations and range in size with each dimension (or each cell) would have been useful for the reader to better understand the results.

Many of the above criticisms and suggestions result from the brevity of the original article. It is quite likely that journal restrictions beyond

the control of the authors contributed to the omission of additional data and tables which would have made the report much clearer for the reader.

INDIVIDUAL STUDIES

Levenson, Hanna and David W. Brooks. "Student Evaluation of Lectures versus Graduate Student Laboratory Instructors in Introductory College Chemistry." Journal of College Science Teaching, 5(2):85-88, 1975.
Descriptors--*College Science; *Chemistry; *Classroom Environment; *Educational Research; Higher Education; Instruction; Science Education; Teaching Assistants; *Teacher Evaluation

Expanded Abstract and Analysis Prepared Especially for I.S.E. by Richard M. Schlenker, Maine Maritime Academy.

Purpose

Levenson and Brooks felt there was a need to investigate the usefulness of student evaluations of laboratory courses because the goals and methods of laboratory teaching differed drastically from those of classroom teaching. They felt the differences in milieus and purpose between the two instructional situations would necessarily affect student perceptions of their instructors in each case; therefore, there would be a differential effect upon student evaluations of instructors depending upon the instructional setting and purpose being considered. It followed, then, that taking into account setting and purpose was of paramount importance when interpreting student ratings of faculty members. Further, the indigent state of knowledge concerning these, the setting and purpose variables suggested a strong need for additional research. In these regards, three primary research questions were confronted.

1. Would students' evaluations of their laboratory instructors differ significantly?
2. Would students' evaluations of their lecture instructors differ significantly?
3. Were there significant differences between students' evaluations of their laboratory instructors and evaluations of their lecturers?

Three working hypotheses were used as guidelines for the investigations. They were: (1) students' evaluations would significantly differentiate among classroom lecturers; (2) students' evaluations would significantly differentiate among laboratory instructors; (3) students would rate their laboratory instructors more positively than their lecturers.

Rationale

While the investigations were not conducted according to a specific model, there were two underlying assumptions which acted as points of departure for the research. The first assumption was that in the absence of sensitive field tested questionnaires, written specifically to evaluate laboratory instruction, educational administrators would be forced to make decisions concerning laboratory study based upon instruments which were invalid for that purpose. Secondly, interpreting evaluations and making subsequent decisions without taking into account the setting might discourage good educational practices and penalize instructors whose primary obligation is to conduct laboratories.

Research Design and Procedures

The research reported herein was conducted at Texas A & M University and the University of Nebraska between the spring semester of 1972 and the spring semester of 1974. It includes one main and two corollary studies. The main study and one corollary study were conducted at Texas A & M University while the second corollary study was conducted at Nebraska University. All subjects participating in the studies were first-year chemistry students.

Samples

There were three samples in the investigation: (1) the main study involved a sample of 329 students randomly selected from a first-year chemistry population of 1,600. This sample was not controlled for sex; however, the authors do indicate the inclusion of insufficient numbers of females in the sample to allow for the control of the sex variable; (2) the first corollary study involved a sample of 200 Texas A & M students; (3) the second corollary study involved a sample of 193 University of Nebraska students.

Methods

The methods used in each study are described by individual study: (1) The main study evaluated professors who spent three hours each week lecturing to students and a short period of time supervising prelaboratory instruction,

and graduate students who spent about three hours each week conducting laboratory sessions but did not lecture formally to the students. During the laboratory sessions the only instruction provided by the laboratory instructor was on a tutorial basis. Following the completion of the spring 1972 semester the subjects were asked to complete two questionnaires (designed by the investigators) which contained items related to most-liked and items related to least-liked teaching techniques. One questionnaire was intended for evaluation of the lecturer while the second dealt with evaluation of the laboratory instructor.

After completion, the questionnaires were content analyzed for both most- and least-appreciated teaching techniques on a 1 to 3 point system. In the former category a score of 1 on an item indicated a high degree of detail and appreciation while a score of 3 indicated little detail and appreciation. The latter category was scored conversely. A score of 3 on an item represented a highly detailed and highly unfavorable rating and a score of 1 represented a favorable rating for that trait but the response lacked detail. Eleven lecturers and ten laboratory instructors were evaluated in this manner.

(2) The first corollary study asked subjects to describe, in detailed writing, the conduct of the last lecture and laboratory session they attended. The scoring of this instrument was based upon word count techniques.

(3) The second corollary study was conducted during the spring 1974 semester and used the same instrument and evaluation scheme as was used in the main study. In this study students evaluated graduate student lecturers who also were their laboratory instructors. The graduate students spent one hour each week as formal recitation instructors and three hours each week as laboratory instructors. They were evaluated as lecturers and as lab instructors in the same manner as were lecturers and lab instructors in the main study.

Data Analysis

All hypotheses were evaluated via analysis of variance techniques and the maximum probability for making a Type I error was set at $P \leq .01$.

Findings

The findings are summarized in two sections. One lumps the findings of the main study and the corollary study conducted at the University of Nebraska and the second includes the findings of the corollary study conducted at Texas A & M University.

Main Study and Nebraska Corollary Study

Although the main study involved both senior faculty members and graduate students while the Nebraska corollary study involved only graduate students, the results of both investigations were essentially the same. They were: (1) questionnaire items concerned with most-appreciated teaching techniques failed to differentiate significantly among lecturers, among laboratory instructors or between lecturers as a group and laboratory instructors as a group; (2) questionnaire items concerned with least-appreciated teaching techniques differentiated significantly among lecturers, among laboratory instructors at the .05 but not the .01 level, and between lecturers as a group and laboratory instructors as a group. In the case of the third finding, subjects were more critical of lecturers than of lab instructors; (3) the most positively rated lecturer was on the positive side of the mean item scores for both most- and least-appreciated teaching techniques and the most negatively rated instructor was on the negative side of the mean for both types of items.

Texas A & M Corollary Study

The following results were obtained from the data analysis in this study: (1) a total word count failed to differentiate significantly between lecture and laboratory settings; (2) subjects wrote more about the first type setting they evaluated; (3) subjects wrote more about lectures if the lecturers were not present and the Ss had just completed a lab session; (4) subjects wrote more about laboratory settings if the lab instructors were not present and the Ss had just finished attending a lecture; (5) when numbers of first person singular and plural pronouns used were analyzed there was a highly significant difference between the laboratory setting and the lecture setting

favoring the lab setting; (6) when lectures were described first, twice as many first person singular and plural pronouns were used than when lectures were described last.

Interpretations

The investigators drew the following conclusions from their findings based upon the use of an open ended or subjective questionnaire format.

1. Least-appreciated teaching technique related questionnaire items can meaningfully discriminate between lecturers.
2. Students are more critical in their evaluations of lecturers than they are of laboratory instructors.
3. Most-appreciated teaching technique related questionnaire items are not good discriminators of lecturers and laboratory instructors.
4. Students preceive lecture and laboratory settings differently.
5. Instruments which are effective in evaluating one educational setting are less effective in making evaluations in another setting.

ABSTRACTOR'S ANALYSIS

In this period of increasing demand on the part of taxpayers and students for faculty accountability, the burden of reconciliation of the issue has fallen upon the educational administrator. Who will be tenured and who will not, which programs survive and which meet their demise and many other difficult questions are related to good and ongoing faculty evaluation programs. They are questions for which there is no simple answer and they are questions with which even the most knowledgeable of administrators finds difficult. It follows that those unskilled in the art of faculty evaluation may serve only to frustrate able faculty members, discourage good programs and generally create havoc amongst stable and good educational environments.

Since many educational administrators enjoy expertise in areas other than faculty evaluation instrument development, testing and validation, the burden of such activities falls upon those so schooled. It is to these concerns and to an obvious void in the literature that the authors have addressed their research.

The concern that instruments developed for the evaluation of one educational setting might not have validity when applied in a different setting is valid. Further, the validity was supported by the results of the study.

The results of this study suggest a need to be extremely cautious when interpreting the results of any faculty evaluation. If the evaluator is not well versed in interpretation and/or familiar with events surrounding an evaluation instrument's development, then advice and counsel should be sought from colleagues and other professionals who are more knowledgeable in the area.

The results of the study suggest that certain types of evaluation items are viable discriminators of weak and strong faculty members while others are not. However, this conceptual contribution must be used with caution since it is based entirely upon student perceptions. To generalize to a more global population at this time would be premature. It might be that, if the research were duplicated at different institutions, controlled for age or sex, the outcome would be somewhat different.

Research design also has a strong influence upon the outcome of a piece of research. It could be hypothesized that different results would surface if graduate student lecturers were evaluated based upon the same number of student-lecturer contact hours per week as the faculty-lecturers. Another hypothesis suggests that if institutional goals and objectives were controlled, the outcome might favor other conclusions. Random sampling for all studies, while perhaps not always essential, always strengthens the validity of conclusions. Generally speaking, when one considers the possibility of institutional constraints, the problem of obtaining willing subjects and the myriad of other confounding variables, only one conclusion can be made concerning the research design; that is that the design was a good one.

Perhaps the most scathing criticisms made of contemporary research in any field are very subtle. Those for whom research would be of most benefit quietly refuse or openly resist reading the latest journal articles. It is not until research is re-written and incorporated in various compendia and other textual materials that many educators become aware of its existence and by this time it is old and perhaps dated. The question why such events come to pass is not an easy one to answer.

Placing the reading and subsequent use of contemporary research within its proper context, it must be remembered that the majority of research is published as journal articles. Journal editors today are being confronted with periodic price increases which seem geometric. Since production costs equate to document and individual article length, editors and editorial advisory boards encourage investigators to force the most information into the smallest space possible. The concomitant results of these events lead to the production of articles understood by only those well versed in the particular subject area.

With these comments in mind, several criticisms and suggestions are advanced which might make this excellent piece of work easier to follow:

1. The use of the phrase "is to make in and between comparisons of the ratings students give to their lecturers and laboratory instructors in introductory chemistry" is vague. Many readers would find the use of "in" and "between" difficult to understand and become frustrated early on in their reading. A more detailed (not verbose) description of the purpose of the paper would aid the reader greatly.

2. In the case of the main study, it is difficult to tell whether the lecturers were being evaluated based upon three hours of lecture each week or upon the supervision of the prelaboratory instruction. A need for clarification exists here.

3. The open-ended evaluation questionnaire would be easier to understand if two sample items were included, one related to most- and the other related to least-appreciated teaching techniques.

4. The reader would be well served if the fact that there was a main and two corollary studies had been mentioned at the beginning of the "Methods" section of the paper.

5. One must question how the subjects were chosen for the two corollary studies. Also, were the 200 subjects in the Texas A & M corollary study some of the same subjects who participated in the main study and/or were they drawn from the same chemistry population as the subjects in the main study?

6. Did the Texas A & M corollary study evaluate the same lecturers and laboratory instructors as did the main study?

7. The results of the main study are located in the "Results" section while those of the two corollary studies are found in the "Discussion" section. The paper would be easier to follow if all of the research results had been included in the "Results" section.

As was previously mentioned, the state of the art concerning the use of evaluation instruments across educational settings can only be described as indigent. However, the results of this research point to a need for additional work in the area, especially if educational administrators, faculty development officers, and others are to use student ratings to best advantage. Therefore, the following suggestions are made for additional research.

1. The sex variable should be investigated. Specifically, does the sex of the student differentially effect the outcome of the rating, does the sex of the lecturer or laboratory instructor differentially effect the outcome of the evaluation, and are there any sex-sex interactions which confound results when sex is used as a variable?

2. Since this research suggests a difference in the relative abilities of certain instruments to evaluate different educational settings, there is a need to continue development of evaluation instruments for specific educational settings.

3. The question of whether subject age is a factor in the way an instrument differentiates within educational settings should be investigated.

4. The question of whether an evaluation instrument developed for use in one laboratory subject area can also be used in other laboratory subject areas needs to be answered.

5. Finally, the question of whether or not this study is duplicable, should and must be answered. It should be remembered that the generalizability of a study's results depends upon the results of similar and identical studies. To generalize based upon the results of one study courts confusion and the propagation of half-truths. In this regard, one might question what effect the study's designers had upon the outcome of the study.

REFERENCES

Levenson, H. and David W. Brooks. "Student Evaluation of Lectures versus Graduate Student Laboratory Instructors in Introductory College Chemistry." Journal of College Science Teaching, 5(2):85-88, 1975.

Za'rour, George I. "Science Misconceptions Among Certain Groups of Students in Lebanon", Journal of Research in Science Teaching 12(4): 385-391, 1975.

Descriptors--*College Science; *Educational Research; Higher Education; Science Education; *Secondary School Science; Secondary Education; *Scientific Concepts.

Expanded Abstract prepared by Eugene D. Gennaro, University of Minnesota.

Purpose

The major purpose outlined by the author in this study was to identify erroneous notions (science misconceptions) about some scientific facts and concepts and to determine the extent to which they are prevalent among certain groups of students in the Beirut, Lebanon area. The study also aimed at determining if proneness to science misconceptions is related to the variables of years of education, sex, major in science, achievement in science, and culture.

Rationale

One of the considerations in effective science teaching, suggests Za'rour, is the identification of misconceptions in the popular notions of scientific facts and concepts. Za'rour reports that Weaver (1965) conducted a survey of physics misconceptions present in twelve series of science textbooks for elementary schools and seems to have found none free of misconceptions. The author also reports that Garone (1960) pointed out that science misconceptions could be traced to improper reliance on common sense and to the misinterpretation of one's experiences. Za'rour suggests that if misconceptions are related to irrational thinking or to a misinterpretation of the cause-and-effect relationship as explained by Hancock (1940), then proper teaching-learning situations aimed at fighting these shortcomings should help in reducing misconceptions. Za'rour reports that Kuethe (1963) has shown that the awareness of the prevalence of science misconceptions on the part of the teacher can direct teaching toward a clear differentiation of a concept from other concepts that have a high probability of intrusion.

Research Design and Procedure

The subjects were high school freshmen and juniors from 11 high schools and practically all university sophomores at the American University of Beirut (A.U.B.). The combined total was 1,444 students. Except for 130 American students from the American Community School (A.C.S.), all the students were, or had been before joining the university, part of the Lebanese system of education. All students in the sample were from a wide variety of schools in Beirut and its surroundings. University students who did not study through the Lebanese system of education in their high schools were excluded.

The process of developing the test was performed through a review of previous studies, interviews with teachers, and drawing from the author's teaching experience. Two tests with a total of 64 items were piloted. These consisted of multiple-choice, true-false, and open-ended items. The idea of a multiple-choice item was retained if it had a distractor which was more attractive to the pilot-study students than was the correct answer. The ideas of true-false and open-ended questions were generally retained if the majority answered them incorrectly. Distractors which were not attractive at all were replaced and erroneous responses to the open-ended questions were transformed into distractors of multiple-choice questions. A new version of the tests was tried to check its language and construction. Final modification resulted in a 40-item multiple choice test with four alternatives per item. About 20 items were in the physics area while the other 20 were distributed among earth and space science, chemistry, and biology. The 120 distractors include 12 of the "none of these" and "impossible to tell" type. If these were set aside, 108 erroneous science statements or potential misconceptions were left.

The test was administered to the students in their respective schools or classes by the researcher or an assistant. The percentage popularity of each alternative was computed for the total sample and for each of the different subgroups which were involved in the variables to be studied. The responses of the 130 American students from A.C.S. were analyzed separately throughout the study. A distractor which was selected by a percentage greater (at the .05 level of significance) than the expected

chance score was labeled a misconception. t-tests were used to test for significance. A Pearson r was calculated between the test score of correct responses on the misconception test and each of the following: science grades of the students from two high schools, and the verbal and quantitative scores on the Scholastic Aptitude Test (SAT) of the eleventh graders from A.C.S.

Findings

The results indicated that 20 out of the 108 distractors or potential misconceptions were selected by 30 percent or more of the respondents. Tables in the report list the misconceptions and their average popularity and percentages at the different class levels. Considering the significance of difference in percentages between the ninth grade and the university group in these items, it was found that there was a significant decrease in nine cases, an increase in one case (the weight of an object at the North Pole when compared to its weight at the equator is smaller) and no significant change in the remaining ten items.

In studying sex differences, there appeared to be a difference in favor of the males in the total number of misconceptions held at the eleventh-grade level. The differences between males and females at the ninth grade and university level was much less pronounced.

Comparing the students of the A.C.S. school with comparable Lebanese students did not reveal appreciable differences. There was a qualitative difference, however, in that there were different items misconceptualized by each group. Students of these two schools did significantly better than schools enrolling students from a lower socioeconomic class.

The results reflected a significant improvement in performance of those students in a university who had or were completing two to four science semester courses at the university when compared to those with no or one science course. For the group of American eleventh grade students who had in their records the SAT scores, the correlation coefficient between the scores of correct responses on the misconception test and each of the

verbal and quantitative SAT part scores were .41 and .68, respectively. At the two schools where science grades were obtained, the correlation coefficient between the scores of correct responses and science grades were .51 in one school and .27 in the other.

Interpretations

A significant steady decrease in adherence to misconceptions with increase in education occurred only in two items which are included in the science curriculum of Lebanese schools, while 4 of the 20 items shown to be insensitive to level of education are not directly taught as part of the curriculum. One item (the most misconceived by most students), "When compared to moist air, the density of dry air is smaller" seems to be due, the author states, to a misconceived common-sense notion that wet objects weigh more than when they are dry and this is then generalized by students to dry and moist air.

The results of this study showed that the females held significantly more misconceptions (11 out of the 20) than did the males at the eleventh grade, but there was little difference between ninth grade girls and boys and university men and women. This appears to be somewhat at variance with findings of Baily (1962) who reported an overall tendency for boys to hold fewer misconceptions than girls and Adler (1966) who found that men college students surpassed women students in the understanding of science concepts.

The qualitative differences, the author states, between the performance of the American students and those of the Lebanese students of comparable status may be due to cultural differences at home and/or different methods of teaching and curricula. Considerably more American students than comparable Lebanese students have the misconception that air is mostly composed of oxygen. The author raises the question: Do the American students speak more of the occurrence of oxygen or are local students drilled on remembering facts about percentages of gases in the air or is it a combination of these two factors?

The finding that fewer misconceptions are held by those who were completing more science courses at the university is in agreement with the results of Adler (1966) and is in disagreement with Boyd (1966).

The study, the author hopes, will motivate teachers to analyze the incorrect responses of students on test questions because of the valuable feedback that it can provide and in the assistance it can give to the teaching-learning situation.

ABSTRACTOR'S ANALYSIS

This piece of research suggests to teachers that they not only look at the right answers that are given by individuals and/or groups in examinations but also be particularly alert to those questions where a significant number of students give a wrong answer to a question. The teacher, then, can counteract the misconception by providing the proper curricular material or experience to correct the misconception.

The use of open-ended questions and then studying the responses is necessary and good but a time-consuming method in discovering the misconceptions students have. The researcher and science teacher need to carefully examine responses on tests and then lump similar answers in determining common errors. The pay-off is that these can be converted into more objective (such as multiple-choice) questions which then can be used to quickly survey a number of areas and to alert teachers and researchers to problem areas. In summary, the design of the test employed in this study seems to be a fruitful way to approach the discovery of students' misconceptions.

The misconceptions in the test reported in this research vary from fact (e.g., "Air is mostly composed of oxygen" and "Bones make up most of the weight of the human body"), to concepts which are basic to certain understandings in science (e.g., "When compared to the work that is put into it, the work can be obtained from a simple machine is usually greater" and "Ships float on water because they are made of material less dense than water"). It is more important, obviously, to identify misconceptions having to do with concepts than those having to do with facts. For the

most part in this study the misconceptions deal with concepts rather than merely facts.

It would be well to identify those misconceptions which could be remedied by demonstrations or, better still, by hands-on experimentation. Unfortunately some of the misconceptions reported in this study are associated with abstract models [e.g., "As you're listening to a radio using electricity at home, the electrons that flow into the radio all change into energy (light, sound, or heat)" and "To change an element such as nitrogen into another element such as oxygen is impossible,"] and no hands-on experimentations can be provided. It is questionable whether expository teaching will easily clear up these types of misconceptions.

The test was composed mostly of concepts from physics. It would be useful to identify the misconceptions that occur in chemistry, biology, and earth and space science. Common misconceptions in biology such as plant cells have cell walls whereas animal cells have cell membranes or plants photosynthesize whereas animals respire are two common biological misconceptions, for example, that the abstractor has experienced in his teaching. If we could identify more of these types of misconceptions made by students, we then might be able to design curricular materials to include sufficient experimentation and material to help alleviate problem areas.

It would be interesting to try Mr. Za'rour's misconception test in American settings at schools varying in socioeconomic levels to find out how widespread and how different are the misconceptions among various groups. Also, cross-cultural studies might reveal under close analysis why it is that students in some cultures experience ease in mastering concepts and why large numbers of students in other cultures are left with erroneous notions.

Also looking at girls and boys in various cultures to see how they differ may give clues to why boys do better than girls in science tests in some societies and not in others. It is not just important to see that there are differences between girls and boys on whole science tests. It is most valuable to know on what items these differences take place. Since the test has more physical science items than items from the life sciences,

it is surprising to the abstractor that significant differences between girls and boys did not occur at the ninth grade and university levels. It may be that, in the Lebanese society, different cultural forces are operating from those in the American culture. It would be well for American researchers to compare differences in boys and girls and men and women on misconception tests in the life sciences; it may be that males and females are on a more equitable footing in the life sciences and less so in the physical sciences, again because of cultural differences.

The National Assessment results in the United States and other standardized science tests can provide additional items where large numbers of students have faulty notions in science. The National Assessment results can be used to identify misconceptions of 9, 13 and 17 year old American students, and these misconceptions added to Mr. Za'rour's list. For instance, for the 13 year old, misconceptions such as the following can be identified by looking at popular choices of the distractors on the test: "When a person faints, one should lay him down and apply cold packs" and "Ice melting most clearly forms molecules different from those present at the start" and further, "A fossil of an ocean fish found in a rock outcrop on a mountain probably means the fossil fish was probably carried to the mountain by a great flood."

Also, it is possible to identify, by looking at the results of this test, what misconceptions are held by 17 year olds. These, then, can be used to add to Mr. Za'rour's list. In addition to the identifying misconceptions in fundamental facts and principles of science, the National Assessment instrument instrument will identify misconceptions about abilities and skills needed to engage in the processes of science and the investigative nature of science. Not only is it important to identify misconceptions having to do with the product element of science but the process element in science as well.

But even after identification of a vast number of such misconceptions in the various scientific disciplines, there remains the difficult task of deciding on the means by which to correct these misconceptions in students' minds. Is it enough to point out the error, give the correct answer and hope the student then assimilates the information? Obviously this will

work with some students, but it may be that in order to correct some basic misconceptions for a great many students, extensive teaching has to be done. Certainly this is true for some of the more basic and difficult concepts embodied in some of the misconceptions.

REFERENCES

Garone, J. E. "Acquiring Knowledge and Attaining Understanding of Children's Scientific Concept Development." Science Education, 44:104-107, 1960.

Hancock, C. H. "An Evaluation of Certain Popular Science Misconceptions." Science Education, 24:208-213, 1940.

Kuethe, L. J. "Science Concepts: A Study of 'Sophisticated' Errors." Science Education, 47:361-364, 1963.

Weaver, A. D. "Misconceptions in Physics Prevalent in Science Textbook Series for Elementary Schools." School Science and Mathematics, 65:231-240, 1965.

Burkman, Ernest. "An Approach to Instructional Design for Massive Classroom Impact." Journal of Research in Science Teaching, 11(1):53-59, 1974.

Descriptors--*Curriculum Development; Individualized Programs; *Instructional Design; Junior High Schools; Program Descriptions; *Science Curriculum; Science Education; Science Materials; *Science Programs; *Secondary School Science

Expanded Abstract and Analysis Prepared Especially for I.S.E. by Robert E. Yager, University of Iowa.

Purpose

The major goal of the report is to describe the goals, rationale, and procedures which were used in developing the Intermediate Science Curriculum Study (ISCS) which was initiated in 1961. The story of the planning and development of ISCS is reported for instructional designers who are concerned with curriculum innovation and student impact.

Rationale

Previous attempts to individualize instruction are mentioned as fairly common occurrences. The ISCS authors began in 1964 with investigations as to why past attempts resulted in little more than local impact. They found that all were the work of few persons with ideas and nearly ideal teaching conditions. All tended to fail when they were transported generally to the real world. This real world involves considering such general issues as school budgets, school design, teacher preparation, parental expectations, and state regulations.

The ISCS developers sought means for considering these problems while proceeding with a practical model for individualizing science for the junior high school years. They argued that solving such problems and developing such a model would take considerable time, resources, money, and full attention of top-level people. They were able to secure a $1,600,000 grant from the Office of Education and later a supplemental grant of $3,500,000 from the National Science Foundation. With such resources, time was available as well as facilities and talented people.

More than 250 persons were involved with the effort, providing a massive flood of ideas and the involvement of persons who would be affected by the program.

Research Design and Procedure

Initial efforts included estimates concerning the degree of innovation that junior high schools could and would tolerate. It was found that teachers were shallowly and narrowly prepared in science. Many were non-majors. The style of teaching was didactic with few activities, especially those that could be called problem-solving. An early decision was to build the desired content and instructional rationale into the materials to enable teachers to use them with little or no special training.

A program for use in rooms with poor laboratory facilities and resources was planned. Low-cost kits were designed which contained all materials needed. Anything requiring special facilities in the classroom or expensive equipment was eliminated. Printed materials were selected as the primary vehicle for communication with students because of state textbook regulations, teacher skills with media, and tradition.

Findings

Two types of instructional materials were produced. Core materials deal with concepts that are important for all students. Optional activities are called excursions and are of two kinds. One kind of excursion provides for remedial work while another provides interesting extensions for motivated students. The materials include about equal amounts of core and excursion materials.

The materials assume that all students will complete all of the core materials but at a pace the student elects. Students encounter "check-ups" within the materials that may direct students to remedial activities if needed. The materials include references to other excursions which

include concepts in new settings, historical development of concepts, or more quantitative consideration of a concept.

The materials purport to utilize the Piagetian model for intellectual development. The developers seek to emphasize the usual psychological transitions which occur in the junior high school in a number of ways. Most activities require students to handle concrete objects. The more abstract concepts are delayed to the end of the sequence when students can better handle some abstraction. Disequilibration is facilitated by continually putting students into problematic situations for which rational explanations are sought. Choice of concepts that are included are based upon whether they are needed and will be reinforced by subsequent use. Questions are included frequently as a means for increasing student motivation and involvement.

Interpretations

The author describes the ISCS approach as a "semi-systems" one. The usual products of a systems approach are specific instructional objectives, instructional materials directed at accomplishing the objectives, and evaluation materials to determine when objectives are met. The procedure actually used, however, is termed "semi-systems" since the objectives were formulated after the development of instructional materials. The ISCS group found that the writing group was unproductive as they grappled with stating specific objectives. When the decision was made to abandon the effort, there was a burst of productivity and a volume of "good" instructional materials resulted. The developers state that scientists often believe that most really important objectives of science instruction can never be stated concretely. The ISCS developers concurred with this rationale since scientists were needed and they could not work productively when asked to prestate instructional objectives. The "semisystems" approach resulted in quick production and school trial.

The activities which occurred between June 1967 and May 1968 were selected as typical for describing the ISCS developmental effort. This included a seventh grade revision and a first draft eighth grade program. About

40 scientists, teachers and "educationists," supported by artists and editors, were responsible for the written materials, equipment kits, and evaluative instruments for 10,000 students to use during the 1967-68 school year.

The project staff processed seventh grade feedback information from the field trial of the preceding year. They also outlined the planned eighth year course. The summer writing teams were then assigned specific tasks with rather rigid time-lines. Each seventh grade team was assigned a specific portion of existing materials to be revised. All seventh grade materials were revised within three weeks. Each eighth grade team was given one week to draft one subject with a rough draft of the entire course completed in ten days. Reconstituted teams then produced a second draft with another one-week deadline. The process was repeated several times with all parts of the program undergoing at least three drafts.

The authors prepared prototypes of all equipment. Suppliers were contracted to assemble the actual apparatus kits. By mid-September all kits and trial editions were in schools in five test centers for field trial. In addition to the actual field testing, a group at Florida State experienced the materials via a computer assisted program. This provided another important vehicle for evaluating the program.

As the evaluation information was collected, formal performance objectives were committed to paper. They were drafted by project staff members. Self-tests were developed quickly after the objectives were formulated. Although large numbers were utilized for the summer writing conferences, the final editions involved relatively few persons. The ISCS experience suggests that many persons are good for mass productions, for ideas, for inventions. However, relatively small numbers are best for refinement and preparation of a final version for publication and widespread use. The ISCS program is described as an innovative one utilizing an individual approach. It has had major impact upon junior high school classrooms.

ABSTRACTOR'S ANALYSIS

"An Approach to Instructional Design for Massive Classroom Impact" is an interesting account of the ISCS story as viewed and reported by the initial director of the project. This kind of analysis is important for the research community. In many respects it is the kind of observational report for which many researchers yearn. Yet it is not a research report in any traditional sense.

The report is a brief account similar to the Grobman book describing the Biological Sciences Curriculum Study (BSCS) effort (1970) and Karplus and Thier's description of the Science Curriculum Improvement Study (SCIS) project (1968). Both of these accounts, however, are more inclusive and provide far greater documentation.

Relating the ISCS effort to other attempts at individualizing instruction is both interesting and valuable. Similarly, terming the effort a "semi-systems approach" provides a valuable framework for assessment and comparison. The analysis of Piaget's model and the ISCS program is also of general interest. It would be of greater interest for the author or others involved with ISCS to develop any one or all of these strands into a comprehensive analysis. In one sense this article merely scratches the surface and does not establish the assertions as such in the program.

The author does not approach science curriculum theory and the part ISCS might play. Certainly there are major differences both in the materials, the times, and schools from the national curriculum efforts in science pre-ISCS and post-ISCS. At times the author seems less intent upon reporting specific approaches that were used several years earlier as opposed to emphasizing procedures and directions for 1974 and beyond.

All too little can be included in a seven-page manuscript of this kind. However, observational information is valuable and should be collected. It would be of interest to compare this 1974 report of the "approach" to ISCS design to reports released in 1967 and 1968. It would also be valuable to reduce this report of instructional design to its simplist procedures and to compare it to other national efforts—both in science

and other fields. Further, it would be worth comparing a more complete description of the ISCS development with the reports of Grobman on BSCS, Karplus and Thier on SCIS, and other directors of current national curriculum efforts. This could give us an interesting 20-year history concerning one of the most extensive curriculum development periods ever.

"An Approach to Instructional Design for Massive Classroom Impact" was well-written and of general value. Because of its brevity, however, there were many assertions, arguments, statements of fact that create unanswered questions. Some of these include the following:

What factors (features) of ISCS design make it "innovative"? What is the history of individualization? How does pacing of programmed materials result in individualization? Which other "individualized" programs were developed with the view of general use? Where is the specific information that describes the 1966 science teachers, junior high schools, the existing materials? What is the average cost for junior high science? Pre-ISCS? Currently? With the information presented (including rationale concerning content and approach for both teachers and students) what is the explanation for the need of special teacher preparation materials? What are some of the specific outcomes of ISCS instruction? What are the objectives and to what degree have they been met? Is "impact" measured only in terms of numbers of students? What is the source of the figures regarding students (from publishers)? What is the "experience that has borne out that interest has tended to remain high"? Is there specific information concerning student motivation with ISCS materials. How were initial staff objectives for materials (and for students) different from those formulated after field testing some ISCS materials? Since objectives were staff prepared anyway, is it valid to report that "scientists found it difficult to prestate objectives" as a reason for delaying them?

Some other research has been suggested including a more extensive report of ISCS development and a comparison of it to other national programs. In addition, the information in this general summary report could be compared to staff memoranda, reports to funding agencies, communications with test center personnel, and reports made at professional meetings. Such comparisons would provide additional information both regarding the design and instruction materials, and information as to impact.

Information from authors, teachers, and other staff members would also be of interest. In fact, differences in perceptions could provide some of the most meaningful information for discussion and analysis. Although information from the initial project director is of great value and interest, it is easy to see how such perceptions could be slanted. This makes the report no less valid, appropriate, nor significant.

Certainly a complete accounting of the development of the ISCS design would be a lengthy report—and probably inappropriate for most journals such as the Journal of Research in Science Teaching. The article, as it appears, is probably the most that could be expected. However, this reviewer would favor a more extensive manuscript—perhaps a book like those cited earlier.

If specific data exist that would provide the background for some of the assertions, this information should be published with the usual facts, tables, and analyses. If such information has been published, it should have been noted in this paper. Such reports would add immeasurably to the field and provide needed information for decisions concerning instruction in science.

REFERENCES

Grobman, Hulda. Developmental Curriculum Projects: Decision Points and Processes. Itasca, Illinois: F. E. Peacock Publishers, 1970.

Karplus, Robert, and Herbert D. Thier. A New Look at Elementary School Science. Chicago: Rand McNally and Company, 1968.

McCurdy, D. W. "An Analysis of Qualities of Self-directedness as Related to Selected Characteristics of I.S.C.S. Students." Science Education, 59(1):5-12, 1975.

Descriptors--*Educational Research; *Individualized Instruction; Science Education; *Student Characteristics; Science Course Improvement Project; Secondary Education; *Secondary School Science; *Student Evaluation

Expanded Abstract and Analysis Prepared Especially for I.S.E. by Gerald G. Neufeld, Brandon University.

Purpose

The study was designed to assess the relationship between students' self ratings of ten self-direction skills and four "independent" variables: success in the Intermediate Science Curriculum Study (ISCS) program, level on the course (Levels I, II, or III), school attended, and sex. The ten self-direction skills were: (1) operating independently, (2) seeking answers without assistance, (3) using class time effectively, (4) planning work, (5) using basic study skills, (6) doing the activities independently, (7) adapting activities and assignments to needs, (8) working at a pace commensurate with perceived ability, (9) using excursions, and (10) collecting their own laboratory materials.

Rationale

Many recently developed science programs, including ISCS, require students to assume an active role in directing their own learning. This study examines two basic questions relating to the use of these programs. First, whether students' self-direction skills improve as they work through the program. And, second, whether achievement in the program is related to the students' self-direction skills.

The study was not designed to test any theoretical model and the author does not cite any of the related studies in the field of learner-controlled instruction.

Research Design and Procedure

The study uses a one-shot case study design. A total of 1108 junior high school students from six schools in the Omaha, Nebraska area served as the subjects. They were not randomly selected but were the students of a group of teachers enrolled in an ISCS inservice course. The 535 grade seven students had about seven months experience with Level I. The 410 grade eight students had all completed Level I and had about seven months experience with Level II. A high, but unspecified, proportion of the 164 grade nine students had not completed either Level I or II and so had only about seven months experience with ISCS, while the remainder had taken ISCS throughout junior high school. The author did not separate out these two groups of grade nine students in the data analyses.

The students' self-direction skills were assessed using an author devised Self-Directed Rating Scale (SDRS). The instrument consisted of ten items, one for each of the skills listed above. The students self rated their own skills by indicating on a five-point scale (supported by three behavior descriptions) the degree to which they perceived they had attained each skill (1 indicated low ability; 5 indicated high ability). A total self-direction score was obtained by adding the ratings assigned to each of the ten items. No indication of the validity or reliability of the instrument was provided.

Student achievement was provided by teacher ratings. Teachers identified students in the top 15 percent and the bottom 15 percent of their classes in terms of grades. This provided three achievement groups: the top 15 percent, the middle 70 percent, and the bottom 15 percent.

The data were analyzed using ANOVA for the "k" group comparisons (achievement, level, and school groups) and the t-test for the two group comparison (sex). For the "k" group comparisons, only the ANOVA F scores are provided.

Findings

Statistically significant differences were found among the high, middle, and low achievers for each of the ten skill items and the total self-directedness score. Except for one item (adapting curriculum), the high achievers

perceived themselves as more self-directed than the middle or low achievers. The author hypothesized that the means for adapting curricula are in the reverse order because high achievers tend to be more conforming and are reluctant to skip activities and/or assignments.

The analysis by levels (time spent using ISCS) indicated non-significant differences except for the item relating to the use of excursions and the total score. In these cases the ratings were highest for Level II students followed by those for Levels III and I. The author hypothesized that the varying experience of the Level III subjects is responsible for the lower scores in Level III.

The analysis by school attended indicated significant differences on the total score and all items except for the one relating to seeking answers independently. The reasons for the differences are not revealed by this study.

The analysis by sex indicated several significant differences. Girls scored higher on the total score and items relating to: using class time, planning a work schedule, using study skills, pacing, using excursions, and collecting lab materials. Boys scored higher on seeking answers independently and adapting curriculum. The author hypothesized that the boys scored higher on these items because society tends to value aggressiveness and independence in males.

Interpretations

The author concluded that: (1) success in programs like ISCS requires adequate self-direction skills, (2) these skills improve with increased experience with the program, (3) school "climate" or ways schools or teachers use the program may affect students' perceived self-directedness, and (4) girls saw themselves as more self-directed than boys.

The implications drawn were: (1) that self-directedness should be assessed early in the school year to provide diagnostic data as a basis for adapting instruction or fostering skill development, (2) that school "climate" can

affect the success of a program like ISCS, and (3) that boys are more likely to have difficulty with these programs than girls.

ABSTRACTOR'S ANALYSIS

This study deals with an important area of educational research—the relationship between measurable student characteristics and the effectiveness of various instructional methods. This area of research has generated a great deal of interest and research over the years as it is the key to effective individualization of instruction. Unfortunately, the author does not cite any of the previous work done nor appear to build on the previous research in independent study, learner-controlled instruction, adapting instruction to student needs, or self evaluation.

The validity of the study is undermined by the choice of research design. The one-shot case study design is one of the weakest possible designs and fails to control for many factors that can jeopardize the internal and external validity of an experiment. The use of the design is particularly inappropriate for a study that attempts to determine whether students' self-direction skills improve as a function of time. A time-series design with a control group would have been much more appropriate.

Using teachers' reports of grades in ISCS as the basis for rating student achievement is somewhat questionable. The Individualized Teacher Preparation module Evaluating and Reporting Progress (ISCS, 1972) encourages teachers to consider such subjective factors as: self-pacing efficiency, self-reliance, and social responsibility in determining a student's grade. The reported relationship between achievement and self-direction skills may simply be evidence that teachers are considering these skills in determining the grades.

One of the most serious weaknesses of the study is the lack of any validity and reliability data on the SDRS instrument. In the absence of such data all of the findings are suspect. The relationship between self-direction and achievement may be just a function of grading practices or just a "halo" effect. The variation among the schools may be more a function of SES than "climate." The higher self-directedness among girls may just be an

expression of the general tendency of junior high age boys to rate themselves lower in achievement than girls do (Russell, 1953). The validity of the study would have been improved if the independence scale of one of the standard personality tests had been used. If none of the published tests were suitable and a new scale had to be created, the author should have provided some indication of the instrument's validity and reliability as Pare and Butzow (1973) and Wang and Stiles (1976) did when they devised similar instruments.

The data obtained from the use of the SDRS represent, at best, ordinal data. Although the use of parametric tests, such as ANOVA and the t-test used by the author, to analyze ordinal data is common in educational research and can, to some extent, be justified by reference to the rigor of these tests, the use of non-parametric methods would have been more appropriate.

The analysis of the relationship between self-directedness and time in the program would have been improved if the Level III students had been separated into two groups: those in their first year in ISCS and those in their third year. This would have allowed comparisons among students with one, two, and three years of experience as well as a comparison between Level I students and Level II students with one year of experience.

The manner of presentation of the results of the statistical tests could also have been improved. Tables I, II, and III list the results of the ANOVA tests for differences among groups of students differing with respect to achievement (three groups), level (three groups), and school (six groups) respectively. The F scores listed in these tables indicate only whether or not any of the differences between pairs of groups are statistically significant but not whether all differences are significant or where the differences lie. The failure to indicate the result of subsequent pairwise comparisons (there is no indication that these were even computed) means that it is impossible for a reader to determine which differences are statistically significant and which are not.

The finding of significant variation in self-direction skills among students in different schools is an interesting one. Although this may be the result of SES or school "climate," it may well be due to the fact that the ISCS teachers vary considerably in their use of the program. Subsequent

investigators would be well advised to make use of the Level of Use model (Loucks, Newlove, and Hall, 1975) or the Level of Implementation model (Neufeld, 1978) to attempt to measure and account for this variation.

Many of the new science programs place a great deal of responsibility on the students for managing their own learning. This study represents one attempt to relate student characteristics and program requirements and outcomes for this instructional approach. However, much research is still needed to determine which students can best benefit from this approach and how to optimize student outcomes.

REFERENCES

I.S.C.S. Evaluating and Reporting Progress. Individualized Teacher Preparation series. Morristown, NJ: Silver Burdett, 1972.

Loucks, S.; B. Newlove; and G. Hall. Measuring Levels of Use of the Innovation: A Manual for Trainers, Interviewers, and Raters. Austin: The Research and Development Center for Teacher Education, The University of Texas, 1975.

Neufeld, G. Assessing the Degree of Implementation of the Important Features of a Curricular Innovation. Paper presented at the Bat-Sheva Seminar on Curriculum Implementation and its Relationship to Curriculum Development in Science. Rehovot/Jerusalem, Israel, July, 1978.

Pare, R. and J. Butzow. "The Reliability and Predictive Validity of a Test of Independence of Work Habits." Educational and Psychological Measurement, 33(4):963-965, 1973.

Russell, D. "What Does Research Say About Self-Evaluation." Journal of Educational Research, 46(8):561-573, 1953.

Wang, M. and B. Stiles. "An Investigation of Children's Concept of Self-Responsibility for Their School Learning." American Educational Research Journal, 13(3):159-179, 1976.